领域内中文科技文献中新发现语言描述特征分析*

毛琛瑜 [1,2]　乐小虬 [1]
[1](中国科学院文献情报中心　北京 100190)
[2](中国科学院大学　北京 100049)

摘要: 【目的】分析领域内中文科技文献新发现语言描述特征。【方法】语义标注新发现语言描述特征, 通过句式分析、频次分布统计以及共现分析探究其特征规律。【结果】总结得到领域内中文科技文献新发现语言的句型, 找出新发现语言的特征搭配。【局限】结果具有领域学科局限性, 需要进一步对比研究。【结论】利用语义标注、频次统计以及共现分析可以有效地发现中文科技文献中新发现语言的描述特征。

关键词: 新发现　语言特征　语义标注
分类号: TP393

1 引　言

科技文献旨在为同一问题的其他研究者提供新知识[1], 称得上科学研究成果的论文, 一定要有新发现、新假设或新理论[2]。科学发现、理论创新等是科技创新的重要体现[3], 因此作者写作时会采用特定描述方式声明其首创性。从自然语言理解的角度, 分析新发现语言的描述特征, 以实现对文献新发现语言模式的揭示, 对基于规则的文献信息抽取的召回率提高具有重要的实际意义。

为了从语言描述上把握中文科技文献新发现的特征, 本文以领域内中文科技文献新发现语言为研究对象, 通过语义标注、词频统计、共现分析等方法对新发现语言描述方式进行分析, 探究了新发现语言的描述模式, 以及特征词、句式搭配等特征, 为进一步构建新发现语言模式提供了基础。

2 科技文献中新发现语言描述特征研究现状

科学发现一方面指做出科学发现的过程, 另一方面指科学发现的结果。本文研究的科技文献中的新发现属于科学发现结果的范畴。钱时惕[4]、李醒民等[5]认为从自然界发现新的事实, 或是在科学研究中提出新的概念、原理、假设、定律、述立新的理论体系都属于科学发现的结果; 邱仁宗[6]指出"科学发现必须是发现过去从不知道的新东西, 其参考系是科学共同体, 并且这种发现原则上是可检验的。还有就是科学发现的结果能够结合进科学知识体系中, 成为科学知识的新一章或其补充"。谭暑生[7]指出科学发现与理论创新是指发现新的科学事实和建立新的科学理论(包括正面肯定的和反面否定的), 而这些新的科学事实和科学理论是对自然界尚未被认识的物质及其特性、物质运动规律和物质新现象的一种揭示和认识, 并且其主要表现形式为学术论文或专著。

以上定义主要针对"科学发现"而言, 在科技文献中体现的新发现内容没有明确定义, 结合前人理论, 本文将科技文献中的"新发现"界定为在本研究领域针对自然现象、事物、原理、特征和规律, 通过研究或者经验, 做出发现和创新, 以及揭示了新事实, 其中

通讯作者: 毛琛瑜, ORCID: 0000-0001-8965-3769, E-mail: maochenyu@mail.las.ac.cn。

*本文系"十二五"国家科技支撑计划子课题"基于文献知识网络的领域学术关系研究与示范"(项目编号: 2011BAH10B06-04)的研究成果之一。

新事实[8]可以概括为揭示业已存在但由于各种原因不为人知的现象或事实、已存在信息有所失真的现象或事实以及揭示第一次出现的现象。

科技文献中新发现语言的描述特征分析属于科技文献篇章分析的范畴，在新发现内容的表达上，一般体现为科技文献的知识声明、作者贡献，或者文献的创新内容等。目前关于科技文献中新发现语言的研究主要有基于语言组织分析的方法，基于修辞学分析的方法以及基于文本挖掘的方法。

(1) 基于语言组织分析的方法主要包括以科技文献中知识声明、创新点为研究对象，对其进行语言学角度分析的一系列研究。知识声明(Knowledge Claim)[9]是指由科技文献作者提出、并且被该学科研究团体所认可的知识增量。新知识声明是指为读者提供新知识的句子，作者会在写作时采用特定的表达声明首创性，如: "We find that…"。新知识声明中包含了新发现的内容,一些研究[10-13]对其进行了分析，在此基础上, Dahl[14-15]进一步研究了语言学、经济学领域研究型论文(Research Articles, RAs)中新知识声明的出现规律，并探讨了新知识声明出现位置与语言学特征的关系，以及与篇章不同部分修辞作用(Rhetorical Function)的相关性，研究结果指出，经济学领域的论文作者为了吸引读者的关注，通常会在表达上使用线索指示词以及惯用表达方式(Signaling Expressions)，如"Our main finding is…"; 在时态上，表达成果或者发现时会采用一般现在时，如"We find…/We argue…"。

科技文献的创新性"体现研究领域最前沿的新颖发现"并且"具有广泛的科学意义"[16]，所以，科技文献中新发现内容包含于创新内容。温有奎等[17-18]从逻辑思维角度以及写作程式角度考虑，分析了科技文献的创新点语言特点，指出了文献中研究目的、理论根据以及研究方法的句型，并从根据创新点特征标记创新句子，进而对其进行抽取。

以上研究对科技文献中涉及新发现的内容进行了语言组织分析，其成果包含特征句型、线索词等，具有一定的借鉴意义。但研究方法为人工分析，没有涉及到大规模的机器学习方法识别等，不利于计算机大规模计算应用。

(2) 基于修辞学的研究主要包括分析文献修辞结构和修辞元话语两方面。Teufel 等[19]将修辞状态(Rhetorical Status)定义为问题结构(Problem Structure)、知识贡献(Intellectual Attribution)等部分，其中"知识贡献"包含了文献中的新发现，利用修辞状态标记文献，实现了对新发现部分的识别。Sandor 等[20-21]定义 CKUs (Claimed Knowledge Updates)为研究论文中承载核心知识点并且总结主要发现的句子，用修辞元话语(Metadiscourse)实现对 CKUs 在文献中的定位。其中修辞元话语是指作者为了使读者更加容易理解而使用的除了"背景"、"方法"、"结论"等小标题之外的具有修辞功能)的元话语，例如"Recent findings allow to get… a picture"。

基于修辞角度处理科技文献文本的方法，重点识别出包含新发现内容的大致区域，即作者贡献部分，并没有具体分析新发现的语言组织以及语言描述上的特征，且主要实验文本为英文，中文科技文献的研究无法直接套用。

(3) 从文本挖掘角度，Ibekwe-Sanjuan 等[22-23]利用修辞和词汇线索词，对语料进行标注，将科技论文的摘要标注为观点(Objective)，新事物(New Things)，结果(Results)，发现(Findings)等部分，人工撰写相关规则，而后通过自动模式生成，对已有的规则进行扩展，实现了摘要的语义结构化，其结果应用于信息检索，如返回"New Things":

[NEWTHINGS] We found that more infrared luminous galaxies tend to have a smaller local galaxy density, being consistent with the picture where luminous IRGs are created by merger-interaction of galaxies….

冷伏海等[24]综合运用语义标注、规则抽取以及正则表达式技术，提出一种面向科技文献的混合语义信息抽取方法，抽取科技文献主要创新研究内容和性能指标，该方法可以迅速从 400 余篇文献全文中抽取出主要研究内容。赖院根[25]提出以创新主题为纽带将期刊论文与专利文献链接研究的框架，结合文本结构信息，采用规则的方法对创新主题进行抽取。

基于文本挖掘的方法对文献中新发现内容进行标注以及抽取，但其规则的指定需要人工参与并借助领域知识库等，人工编撰的规则不能全面地覆盖语言现象，并且抽取的结果在整合中会有噪音数据。

综合以上研究，本文以中文科技文献中的新发现语言为研究对象，利用标注的方式对语料进行特征标

注，结合特征搭配抽取以及共现分析等方法，对新发现语言描述规律性特征进行分析。

3　新发现语言描述特征分析方法

科技文献在表达新发现时有一定的语言学规律，其出现的位置、常用表达、以及特征词、线索词都能表征新发现的语言描述特点。本文采用语义标注的方法，标记出科技文献中新发现的语言特征，进而对特征进行搭配抽取、词频统计、共现分析，以探索新发现语言的描述特征规律。

3.1　文档集特征标注

通过对一定量的科技文献进行人工阅读分析，发现科技文献中摘要、引言、结论等部分可以集中体现文章的新发现内容。选取非结构式科技文献摘要作为实验语料，结合本文对新发现的定义，对摘要中描述新事物、现象、特征、规律等新发现内容的句群进行语言分析，人工标注了其特征词(如“发现”)、短语(如“揭示了…规律”)、句式(如转折句)、结构(如并列结构:“研究发现: ①…; ②…; ③…”)等。标注语料示例如下。

示例 1: 该研究以浙江天童木本植物为对象，通过对小枝大小(横截面积)与数量(稠密度)关系的 #研究#FT #发现#FT :1)小枝稠密度与枝截面积 #显著#FT #负相关#FT …，小枝稠密度在两种生活型间#无#FT#显著#FT#差异#FT; …。

——许月等《浙江天童木本植物小枝的“大小–数量”权衡》

示例 2: …系统 #研究#FT 马尾松家系对不同类型低 P 胁迫的适应机制和 P 效率变异 #规律#FT。#结果表明#FT，参试马尾松家系的苗高、地径和生物量等 P 效率指标均 #表现出#FT #显著#FT 的家系变异，…。

——杨青等《异质低磷胁迫下马尾松家系根构型和磷效率的遗传变异》

3.2　新发现语言搭配特征抽取

语料标注完成后，针对新发现语言的特征搭配进行分析。关于词语搭配, Choueka 等[26]、Benson 等[27]、Church 等[28]认为搭配是重复出现的、具有互相关联性且一定任意性的词的组合。搭配也有狭义和广义之分[29-30]，狭义搭配专指固定搭配，要求另一个词伴随而产生的词汇之间的限制性共现关系；广义搭配指同时出现在上下文中，句法及词汇上有所关联的词汇之间的共现。本文对新发现语言特征搭配界定为，出现在新发现句子的上下文中，并具有一定语法关系的特征词语组合。

在标注中发现，大部分新发现语言的特征搭配都出现在一定上下文中，所以在抽取特征搭配时遍历文摘的句子，从子句中将已标注好的特征词组合抽取出来。如“研究…规律”、“表现出显著…指标”等。

抽取完成后，对所有的特征搭配进行系统归类，如，将提示新发现类别的搭配(“探讨了…特征”、“揭示了…规律”等)归类为“Type”，将提示具体新发现结果的词(“研究发现”、“结果表明”等)归类为“Result”等。具体归类如表 1 所示:

表 1　特征搭配归类对应表(部分)

具体搭配	归类后
提示新发现类别类(探讨了…响应/分析了…特征/……)	Type
提示结果部分(结果表明/显示/研究显示/……)	Result
发现…/研究发现/分析发现/…	发现
显著　增加/高于/低于/…	显著 V
显著　正相关/关系/…	显著 N
最高/低/大/佳/…	最…
随(着)…升高　而　增加/…	随…
趋向于…/趋于…	趋势
……	……

有些搭配虽然符合上述归类规则，但是因为其出现的频次在归类中较高，故单独列出来分析，比如“显著差异”并没有归到“显著 N”中，而是单独统计的。

3.3　特征搭配频次分布统计分析

通过特征搭配频次分布统计分析可以有效发现语料中哪些搭配出现的频率更高些，哪些词语在新发现摘要中分布更加广泛的特征。

特征搭配的 IDF，可以由总文档数目除以包含该搭配词语的文档的数目，再将得到的商取对数得到:

$$IDF = \log \frac{D}{Num + 1}$$

其中，D 是语料中文摘的篇数;Num 是包含搭配的文摘数目，如果该词语不在语料库中，就会导致分母为零，故使用 Num+1。

$$TF = Num_p$$

Num_p 为该搭配在语料集中出现的总次数。

IDF 值衡量的是搭配在整个语料中的区分度，如果一对搭配在某篇文档中出现很多次，但在整个语料中出现的次数并不多，则说明此搭配对这篇文档的主题区分度大；反之，如果某对搭配在单篇文档和整个

语料中出现的次数都很多，则此搭配对文档的主题区分度不大。本文意图识别出可以标识新发现语言特征的共性搭配，所以希望搭配的 IDF 越小越好。

3.4 特征搭配的共现分析

通过共现分析可以发现在新发现语料中哪些特征搭配会经常出现，哪些特征搭配组合具有上下链接等特征。计算方法为，将语料中单篇摘要表示为特征搭配的组合，统计计算两两特征搭配组合的共现次数。

4 实验结果分析与验证

4.1 实验数据来源以及预处理

从文献来源上考虑，影响因子较高、被重要数据库收录的核心期刊中刊载文献包含的新发现内容较多且质量较高；从领域的角度考虑，多个领域对比分析有助于得出普适性的特征规律。本文选取植物、物理、化学三个领域做对比分析，其中，植物学领域选择的科技期刊为《植物生态学报》，化学领域选择的科技期刊为《化学学报》、《高分子学报》、《有机化学》，物理学领域选择的科技期刊为《物理学报》。各个期刊数据从 CNKI[31]上定制导出为 EndNote 格式，后存入数据库，数据表主要字段如图 1 所示：

Index	Name
1	UUID
2	AUTHOR
3	YEAR
4	JOURNAL
5	KEYWORDS
6	CALLNUMBER
7	TITLE
8	ABSTRACTS
9	ISBN
10	AUTHORADDRESS

图 1 初始文摘字段展示

实验中，植物学领域选取前 120 篇文摘，手工标注其中的新发现特征；作为对比领域，化学领域和物理领域随机选择 300 篇进行人工判别标注，分别得到包含新发现特征的文摘 116 篇、114 篇。

对已标注的文摘进行特征抽取，得到各个领域的新发现特征搭配表，随后按 3.2 节中提到的规范方法，对特征搭配进行规范，规范后数据如图 2 所示。

RecNo	id	uuid	phrase
Click here to define a filter			
1	1	0f986d4f8a5c4030943a6e1aed193a1c	#Result#FT
2	2	0f986d4f8a5c4030943a6e1aed193a1c	#明显#FT#V#FT
3	3	0b6089a5b06c4ecfade74c99801124a1	#Type#FT
4	4	0b6089a5b06c4ecfade74c99801124a1	#发现#FT
5	5	09d9f100bdc841ce9e3db005e9b924ea	#发现#FT
6	6	09d9f100bdc841ce9e3db005e9b924ea	#最#FT
7	7	082f6abe4d104b2586678308dce588c6	#Type#FT
8	8	082f6abe4d104b2586678308dce588c6	#发现#FT
9	9	01bd0cfd6f3b4153976abebe7e990e28	#Type#FT
10	10	01bd0cfd6f3b4153976abebe7e990e28	#Result#FT

图 2 新发现特征搭配规范表(截取)

4.2 实验结果分析

实验结果主要包含中文科技文献中新发现语言的主要表达句型、高频特征搭配分析以及共现分析。

(1) 新发现语言的表达句型分析

对新发现语料标注完成后，分别对每个领域的新发现类别进行统计。大体上可分为“影响”、“特征”、“规律”、“关系”等类别。但不同的领域间有差别，具体类别展示如表 2 所示：

表 2 领域新发现主要类别统计

ID	植物学		化学		物理学	
	类别	比例	类别	比例	类别	比例
1	“影响”	23.3%	“影响”	25.9%	“影响”	27.1%
2	“特征”	20%	“性质”	15.5%	“性质”	13.1%
3	“关系”	13.3%	“规律”	6.0%	“规律”	5.2%
4	“规律”	8.3%	“行为”	6.0%	“行为”	5.2%
5	“响应”	8.3%	“结构”	5.2%	“过程”	5.2%
6	“变化”	6.7%	“机理”	5.2%	“分布”	3.5%
7	“差异”	4.2%	“变化”	3.5%	“关系”	3.5%
8	“原因”	3.3%	“作用”	3.5%	“响(效)应”	3.5%
9			“特征”	2.6%	“原因”	1.8%
10			“反应”	2.6%	“机理”	1.8%
11			“原因”	1.2%		

通过分析表 2，可得：

①新发现的类别分布在不同领域内有所不同；

②植物、物理、化学三个领域中，新发现的主要类别都是“影响”、“特征/性质”、“规律”这几方面；

③相对于植物领域，物理、化学领域的类别较为相似，都具有描述机理、行为、以及各种特性/性质的类别，如活性、稳定性、导电性等；并且物理、化学领域的新发现类别更分散。

通过对主要类别的新发现语言进行句型分析，对句型句式有宏观的了解概括，这有利于后期语言正则表达式的设计建立。

主要类别的常见描述句型、短语、词如表 3 所示，其中提示句表示提示新发现类别的句子，结果句包含了具体的新发现结果。结果句一般会有提示词,如“结果表明/发现”。

表 3　“新发现”类别主要句型

类	主要句型、短语(提示句)	主要句型、短语、词(结果句)	
“影响”	研究(了)…对…的影响 观察了…对…的影响 ……	结果表明/显示/发现	…对…有显著(较大)影响 …受到…的影响 显著提高/降低/增加/升高 ……
“性质”(活性、稳定性等)	研究了(光学/电学…)性质 探讨…稳定性 ……	结果表明/显示/发现	最大/佳/优…… 有…明显分界 ……
“规律”	研究(了)…规律 揭示(了)…规律 ……	结果表明/显示/发现	呈显著(正/负)相关 表现出显著… ……
“特征”	对…特征进行了研究 测定了…特征 ……	结果表明/显示/发现	最高/低　最大值 差异显著 ……
“行为”	研究了…行为 观察了…行为 ……	结果表明/显示/发现	…行为更加显著 随着…，而… ……
“原因”	研究了变化原因 探讨了…因素 ……	结果表明/显示/发现	…规律明显 趋势一致 ……

(2) 新发现特征搭配的频次统计分析

对新发现语料中标记的特征搭配进行规范归类后，进行特征搭配的 IDF 计算，并降序排列，每种领域分别选择 Top15，具体如表 4–表 6 所示。

表 4　植物领域特征搭配 IDF 计算

ID	特征搭配	IDF	TF
1	Type	0.0772916743016465	126
2	Result	0.161061557367105	108
3	显著 V	0.903970247486114	90
4	最	1.26943002098058	55
5	随	1.46358603542154	40
6	显著 N	2.02320182335696	28
7	呈 趋势	2.08774034449453	18
8	显著 差异	2.2308411881352	12
9	呈 相关	2.39789527279837	17
10	趋势	2.4932054526027	10
11	呈 关系	2.4932054526027	10
12	差异 显著	2.4932054526027	11
13	发现	2.59856596826052	8
14	N 显著	3.00403107636869	6
15	表现出 趋势	3.18635263316264	4

表 5　化学领域特征搭配 IDF 计算

ID	特征搭配	IDF	TF
1	Type	0.299242894852857	101
2	Result	0.676052747200645	63
3	发现	0.861769892995738	55
4	随	1.53471436623816	30
5	最	2.04553999000415	15
6	明显 V	2.80768004205105	6
7	显著 V	2.80768004205105	7
8	有利(助)于	3.14415227867226	4
9	更	3.14415227867226	6
10	有…影响	3.14415227867226	4
11	明显 N	3.36729582998647	3
12	有效 V	3.36729582998647	3
13	先…后…	3.36729582998647	3
14	较好	3.65497790243825	2
15	较高	3.65497790243825	2

综合三个领域的新发现特征搭配 IDF 统计，可以发现:

①Type 与 Result 类搭配在三个领域中都普遍存在，表明领域内科技文献都会采用新发现类别提示词(Type)与具体

表 6 物理领域特征搭配 IDF 计算

ID	特征搭配	IDF	TF
1	Type	0.258861633916289	96
2	Result	0.54654370636807	71
3	发现	0.747214401830221	64
4	随	1.12528053575027	49
5	明显 N	2.33830317559612	10
6	最	2.33830317559612	14
7	显著 V	2.43361335540045	9
8	明显 V	2.53897387105828	10
9	明显	2.53897387105828	8
10	越	2.65675690671466	14
11	更	2.94443897916644	8
12	先…后…	2.94443897916644	5
13	呈	3.3499040872746	3
14	呈 趋势	3.3499040872746	4
15	显著 N	3.3499040872746	3

新发现结果提示词(Result)结合的形式描述新发现内容;

②"发现"类短语(如"研究发现"、"分析发现"、"比较发现"…)其出现具有一定的领域特征，在化学、物理领域出现较多，对比之下在植物学领域出现较少;

③程度类修饰词如"显著"、"明显"、"最…"等在三个领域内都具有较高的频次，提示了"明显,引人注目的"事实，常用于对照实验的结果表达，是新发现中很重要的一方面;

④在描述某一类规律时，常采用"随(着)…"类的搭配句型，在三个领域内普遍存在。

(3) 新发现特征搭配共现统计分析

根据计算 IDF 时的规范表，计算不同领域特征搭配的两两共现关系。根据高频的共现关系来分析在描述新发现内容时常采用的语言模式，并探究不同领域间模式的异同。计算结果如表 7–表 9 (Top10)所示。

表 7 植物领域新发现特征搭配共现关系统计

ID	nt1	nt2	Occurrence
1	Result	Type	95
2	显著 V	Type	45
3	Result	显著 V	40
4	最	Type	32
5	最	Result	29
6	随 T	Type	24
7	Result	随	22
8	呈 趋势	Result	13
9	呈 趋势	Type	13
10	最	随	13

表 8 化学领域新发现特征搭配共现关系统计

ID	nt1	nt2	Occurrence
1	Result	Type	50
2	发现	Type	31
3	随	Type	17
4	Result	随	16
5	最	Type	12
6	最	发现	8
7	最	Result	7
8	显著 V	Type	6
9	随	发现	6
10	Result	明显 V	5

表 9 物理领域新发现特征搭配共现关系统计

ID	nt1	nt2	Occurrence
1	Result	Type	58
2	发现	Type	35
3	随	Type	26
4	Result	随	22
5	随	发现	13
6	Result	发现	10
7	最	Type	8
8	Result	明显 V	7
9	明显 N	Type	7
10	显著 V	Type	7

通过分析三个领域中新发现特征搭配的共现表，可以得出:

①新发现类别(Type)、结果部分提示词(Result),二者共现的频次较高，说明三个领域内都倾向于采用这种 Type-Result 的形式描述新发现内容;

②"发现"类的搭配句式在物理、化学领域出现频次高，用来引出新发现内容;

③描述具体的新发现时，趋向于用引人注意的词汇，如"显著"、"最…"等。

4.3 实验结果验证

针对 4.2 节中得到的新发现语言表达特征搭配模式，总结新发现类型表达表、新发现结果提示词表、新发现特征词表，设计实验对特征的准确性进行验证。

实验思路是针对一篇测试文摘，首先进行分句，然后按照新发现类别规则、新发现结果提示词、新发现内容特征规则的顺序进行正则匹配，如果该文献中包含有新发现类别搭配、新发现内容特征搭配则判定为新发现相关(结果提示词不是必须选项)。结果统计

其准确率(Precision)和召回率(Recall)。计算公式如下:

$$\text{Precision} = \frac{\text{返回结果中相关文档的数目}}{\text{返回结果的数目}}$$

$$\text{Recall} = \frac{\text{返回结果中相关文档的数目}}{\text{所有相关文档的数目}}$$

测试语料采用与 4.1 节同源的期刊数据,每个领域随机选择 100 篇新的文摘,人工判定是否含有新发现内容,如表 10 所示:

表 10 测试语料中新发现内容分布

领域 / 统计	植物	化学	物理
含有新发现内容(篇)	56	31	50
未含新发现内容(篇)	44	69	50
总计(篇)	100	100	100

实验结果统计如表 11 所示:

表 11 各领域新发现特征识别判定

领域 / 结果	植物	化学	物理
Precision	81.48%	70.00%	62.29%
Recall	78.57%	67.74%	76.00%

从以上结果可以看出,本文总结分析的新发现特征集,其描述新发现内容时,具有较高的准确率和召回率。其中,植物领域的识别效果最好,因为植物领域中新发现的类别较为集中,而化学、物理领域则相对分散,见表 2。综上所述,本文总结的新发现语言特征集具有一定的准确性。

5 结 语

本文以领域内中文科技文献新发现语言为研究对象,通过语义标注、词频统计、共现分析等方法,对新发现语言的描述模式做了初步探索,分析了新发现语言句式、特征搭配等特征,并对不同领域间新发现语言描述的句式、特征搭配等进行对比研究,对其表达上的异同进行分析,实现了对不同领域内中文科技文献新发现语言表达的量化研究。

本文的不足之处在于对科技文献中新发现语言的描述目前只限定于自然科学范畴的文献,并且大量的语义标注工作都是人工进行,较为耗时耗力。需要对表达的模式进行机器学习,进而进行大范围计算。

后续研究中,将以现有结果为基础,探索新发现语言的描述特点,建立新发现语言描述模型,使结果更具有实际意义。

参考文献:

[1] Berkenkotter C, Huckin T N. Genre Knowledge in Disciplinary Communication: Cognition/Culture/Power [M]. Lawrence Erlbaum Associates, Inc, 1995.
[2] 温有奎,吴广印. 碎片化科研创新点动态挖掘研究[J]. 数字图书馆论坛, 2014(7): 25-32. (Wen Youkui, Wu Guangyin. Dynamic Mining of Fragmented Scientific Research Innovation Points[J]. Digital Library Forum, 2014(7): 25-32.)
[3] 朱大明. 科技期刊论文创新性鉴审的四个基本要素[J]. 科技管理研究, 2011(9): 199-201. (Zhu Daming. Four Essential Factors in Appraising Innovation of Papers of Sci-tech Periodicals[J]. Science and Technology Management Research, 2011(9): 199-201.)
[4] 钱时惕. 关于科学发现的思维结构[J]. 科学技术与辩证法, 1989(3): 37-40. (Qian Shiti. Thinking About the Structure of Scientific Discovery [J]. Science Technology and Dialectics, 1989(3): 37-40.)
[5] 李醒民,宋德生,王身立. 思想领域中最高的音乐神韵——科学发现个例分析[M]. 长沙: 湖南科学技术出版社, 1988. (Li Xingmin, Song Desheng, Wang Shenli. The Highest Musical Charm in Ideological Field: Case Studies of Scientific Discovery [M]. Changsha: Hunan Science and Technology Press,1998.)
[6] 邱仁宗. 成功之路: 科学发现的模式[M]. 北京: 人民出版社, 1987. (Qiu Renzong. Road to Success: Scientific Discovery Mode [M]. Beijing: People's Press, 1987.)
[7] 谭暑生. 科学发现与理论创新成果评价标准[J]. 发明与创新, 2006(1): 38-39. (Tan Shusheng. Evaluation Standard of Scientific Discovery and Theoretical Innovations [J]. Invention & Innovation, 2006(1):38-39.)
[8] 周露阳. 论审评学术论文创新因素的指标体系[J]. 编辑学报, 2006, 18(1): 68-70. (Zhou Luyang. Index System for Identifying Innovation Factors in Academic Papers [J]. Acta Editologica, 2006, 18(1): 68-70.)
[9] Swales J M. Genre Analysis: English in Academic and Research Settings [M]. Cambridge University Press, 1990.
[10] Swales J M. Research Genres: Exploration and Applications [M]. Cambridge University Press, 2004.
[11] Hyland K. Disciplinary Discourses: Social Interactions in Academic Writing [M]. University of Michigan Press, 2004.
[12] Hyland K. Metadiscourse [M]. John Wiley & Sons, Inc., 2005.
[13] Hunston S. Professional Conflict: Disagreement in Academic

Discourse [A]. //Text and Technology: In Honour of John Sinclair [M]. John Benjamins Publishing Company, 1993: 115-134.

[14] Dahl T. Contributing to the Academic Conversation: A Study of New Knowledge Claims in Economics and Linguistics [J]. Journal of Pragmatics, 2008, 40(7): 1184-1201.

[15] Dahl T. The Linguistic Representation of Rhetorical Function: A Study of How Economists Present Their Knowledge Claims [J]. Written Communication, 2009, 26(4): 370-391.

[16] Publishing with Us [EB/OL]. [2015-11-13]. http://www.sciencemag.org/site/help/authors/publishing.xhtml.

[17] 温有奎, 温浩, 徐端颐, 等. 基于创新点的知识元挖掘[J]. 情报学报, 2005, 24(6): 663-670. (Wen Youkui, Wen Hao, Xu Duanyi, et al. Knowledge Element Mining in Knowledge Management [J]. Journal of the China Society for Scientific and Technical Information, 2005, 24(6): 663-670.)

[18] 温有奎, 温浩. 关键词与创新点词句群分布分析[J]. 情报学报, 2007, 26(1): 50-55. (Wen Youkui, Wen Hao. Sentence Group Distribution of Keywords and Innovation Idea Words [J]. Journal of the China Society for Scientific and Technical Information, 2007, 26(1): 50-55.)

[19] Teufel S, Moens M. Summarizing Scientific Articles: Experiments with Relevance and Rhetorical Status [J]. Computational Linguistics, 2002, 28(4): 409-445.

[20] Sandor Á. Modeling Metadiscourse Conveying the Author's Rhetorical Strategy in Biomedical Research Abstracts [J]. Revue Française de Linguistique Appliquée, 2007, 12(2): 97-108.

[21] Sandor Á, De Waard A. Identifying Claimed Knowledge Updates in Biomedical Research Articles [C]. In: Proceedings of the Workshop on Detecting Structure in Scholarly Discourse. Association for Computational Linguistics, 2012: 10-17.

[22] Ibekwe-Saniuan F, Chen C, Pinho R. Identifying Strategic Information from Scientific Articles Through Sentence Classification [C]. In: Proceedings of the 6th International Language Resources and Evaluation (LREC'08), Marrakech, Morocco.2008.

[23] Ibekwe-Sanjuan F, Silvia F, Eric S, et al. Annotation of Scientific Summaries for Information Retrieval [C]. In: Proceedings of ECIR'08 Workshop on: Exploiting Semantic Annotations for Information Retrieval. 2008.

[24] 冷伏海, 白如江, 祝清松. 面向科技文献的混合语义信息抽取方法研究[J]. 图书情报工作, 2013, 57(11): 112-119. (Leng Fuhai, Bai Rujiang, Zhu Qingsong. A Hybrid Semantic Information Extraction Method for Scientific Research Papers [J]. Library and Information Service, 2013, 57(11): 112-119.)

[25] 赖院根. 期刊论文与专利文献的链接研究[J]. 图书情报知识, 2011(1): 63-68. (Lai Yuangen. Research on Linking Method Between Periodical Thesis and Patent Literature[J]. Document, Information & Knowledge, 2011(1): 63-68.)

[26] Choueka Y, Klein T, Neuwitz E. Automatic Retrieval of Frequent Idiomatic and Collocational Expressions in a Large Corpus [J]. Journal of the Association for Literary and Linguistic Computing, 1983, 4(1): 34-38.

[27] Benson M, Benson E, Ilson R F. The BBI Combinatory Dictionary of English: A Guide to Word Combinations[M]. John Benjamins Publishing Company, 1986.

[28] Church K W, Hanks P. Word Association Norms, Mutual Information, and Lexicography [C]. In: Proceedings of the 27th Annual Meeting of the Association for Computational Linguistics. 1989: 76-83.

[29] 陈雅菊. 现代汉语词语搭配的自动抽取方法[D]. 上海: 华东师范大学, 2006. (Chen Yaju. Automatic Extraction of Chinese Collocation [D]. Shanghai: East China Normal University, 2006.)

[30] 申修瑛. 现代汉语词语搭配研究[D]. 上海: 复旦大学, 2007. (Shen Xiuying. A Study of Chinese Collocation [D]. Shanghai: Fudan University, 2007.)

[31] CNKI [EB/OL]. [2015-11-05]. http://www.cnki.net/.

作者贡献声明:

毛琛瑜: 设计并实施技术方案、技术路线, 数据采集、数据清洗, 实验的分析, 论文撰写及最终版本修订;
乐小虬: 提出研究方向和主要研究思路, 设计研究方案及技术路线, 文章部分修改。

利益冲突声明:

所有作者声明不存在利益冲突关系。

支撑数据:

支撑数据由作者自存储, E-mail: maochenyu@mail.las.ac.cn。

[1] 毛琛瑜, 乐小虬. nf_idf.xlsx. 植物、化学、物理领域新发现特征搭配 tf-idf 统计(频次大于 1).

[2] 毛琛瑜, 乐小虬. nf_occurrence.xlsx. 植物、化学、物理领域新发现特征搭配共现关系(频次大于 1).

收稿日期: 2015-11-26
收修改稿日期: 2016-03-07

Linguistic Features of New Findings in Chinese Scientific Papers

Mao Chenyu[1,2]　Le Xiaoqiu[1]
[1](National Science Library, Chinese Academy of Sciences, Beijing 100190, China)
[2](University of Chinese Academy of Sciences, Beijing 100049, China)

Abstract: [**Objective**] To analyse the linguistic features of new findings discussed by the scientific research papers in Chinese. [**Methods**] We first annotated these features and then explore their patterns with the help of sentence analysis, frequency statistics and co-occurrence analysis technologies. [**Results**] We summarized the sentence patterns and features of words/phrases for new findings listed by the Chinese scientific articles. [**Limitations**] We only examined papers from the field of natural sciences. More comparative research is needed to analyze papers from other areas. [**Conclusions**] Annotating corpus, counting frequency distribution statistics and analyzing of co-occurrence could effectively identify new findings from Chinese scientific articles.

Keywords: New finding　Linguistic feature　Semantic annotation

Ex Libris 获得 ISO 27018 云隐私认证

Ex Libris 于近日获得 ISO/IEC 27018:2014 证书，这是一个由国际标准化组织(ISO)于近期颁布的国际标准协议，旨在提供云计算服务领域中对于个人身份信息(Personally Identifiable Information, PII)的保护指南。ISO/IEC 27018 标准建立了被普遍接受的控制目标、控制对象和行为准则，以确保 PII 数据被云计算服务提供商处理时得到了适当的保护，为云计算服务提供商提供了一个普遍的行为框架。这一证书给予 Ex Libris 所服务的客户以足够的信心，即 Ex Libris 能在云计算中以最高级别来保护个人身份信息。

想要获得 ISO/IEC 27018:2014 证书，公司必须展现出持续不断的结构化措施来保护个人身份信息和用户数据。通过与 ISO 27018 相契合的过程, Ex Libris 表明 Ex Libris 环境能保护个人信息，这种保护遵循了数据隐私法律，允许客户保留对他们个人信息的完全控制，同时客户的数据不以任何非官方目的被使用，另外公司在客户的数据如何被储存和使用方面是完全公开透明的。

Ex Libris 致力于为用户提供高度安全和可信赖环境来进行基于云计算的 SaaS 应用。Ex Libris 已经开发出覆盖云计算服务各个方面的多个层次的安全模型。这一安全模型和控制做法基于国际化协议、标准和工业最佳实践，包括 ISO/IEC 27001:2013、ISO/IEC 27018:2014 和 CSA Star Self-Assessment。

“每一个 Ex Libris 的人都很自豪 Ex Libris 在世界各地的操作平台和数据中心能够获得极为重要的 ISO/IEC 27018:2014 证书。” Ex Libris 的隐私和监管官 Ellen Amsel 说道。“作为本领域基于云计算的 SaaS 解决方案的先锋者，我们采用‘最高栅栏’的做法来贯彻最为严格的国际信息安全标准。我们相信隐私保护是正确应当的，而非不切实际的要求，并且保护我们用户的数据和隐私是至关重要的。”

Ex Libris 的信息安全官 Tomer Shemesh 补充: “一直以来，我们都采用国际标准, ISO/IEC 27018:2014 证书(的获得)是我们将客户利益摆在首位的一个最新例子。遵从国际标准需要在处理 PII 方面有较高程度的经验，并投入较多的时间和资源花费。这一 ISO 证书能使客户对存储在我们数据中心的客户信息的安全和隐私保持绝对放心，并且能使他们以无可匹敌的可见性、规范性和信息安全性来使用我们的云计算服务。”

(编译自: http://www.librarytechnology.org/news/pr.pl?id=21591)

(本刊讯)